DOCUMENT RESUME

ED 334 245 TM 016 797

AUTHOR Trevisan, Michael S.
TITLE Reliability of Performance Assessments: Let's Make Sure We Account for the Errors.
PUB DATE Apr 91
NOTE 35p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education and the National Association of Test Directors (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Computer Software; *Educational Assessment; Estimation (Mathematics); *Performance; Psychometrics; Standardized Tests; *Test Reliability
IDENTIFIERS *Intraclass Correlation; *Performance Based Evaluation

ABSTRACT

Some of the issues regarding the estimation of reliability for performance assessments are explored, and a methodology is suggested for determination of reliability. In performance assessment, the magnitude of the reliability coefficient may be less than that obtained for a standardized test, since part of the potential use of performance assessment is to obtain outcome data for targets difficult to assess with standardized tests. A further problem is the lack of training of many assessment specialists with regard to performance assessment. In addition, reliability may refer to more than internal consistency. There is a growing consensus that the intraclass correlation coefficient (ICC) provides the best means for establishing reliability for quantitative observation data. Three basic ICC models are discussed: (1) one-way analysis of variance; (2) two-way random effects; and (3) a two-way mixed model. The use of ICC methods is discussed, and computer softwares for ICC analyses are described. Four tables illustrate the discussion, and a 26-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334245

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
MICHAEL S. TREVISAN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Reliability of Performance Assessments: Let's make sure we account for the errors

by

Michael S. Trevisan

Multnomah Education Service District, Portland, Oregon

Paper presented at a symposium entitled **Measurement Issues in Performance Assessment**, held at the Annual Meeting of the National Council on Measurement in Education and the National Association of Test Directors, Chicago, IL, April, 1991

TM016797

BEST COPY AVAILABLE

Performance assessments are becoming part of the achievement data routinely collected by school districts and state departments throughout the country. Because of the nature of the scoring, this type of assessment provides unique situations for the assessment or evaluation specialist charged with providing evidence for the dependability of the measures. Reliability and measurement error must be defined somewhat differently than is typical for machine-scored standardized achievement tests. A major source of error in performance assessments for example, is differences due to rater judgment; therefore, a reliability coefficient must often take this error into account. This paper explores some of the issues regarding the estimation of reliability for performance assessments and provides appropriate methodology.

Concern for the reliability of data from performance assessments and the dearth of existing reliability information has been voiced (Rothman, 1990; Suen, 1991). Caution, therefore, has been recommended before wholesale acceptance of this type of assessment is given. The psychometric quality of large-scale performance assessments conducted at the district or

state level will in part be determined from the reliability data. However, how best to assess the reliability of performance assessment measures continues to be a considerable point of confusion among many practitioners (Cone, 1977; Suen & Ary, 1989). In short, performance assessments are trapped on one side by those voicing concern over the lack of reliability data and boxed in from the other side by those uncertain about how to properly assess the reliability of quantitative observation data.

Expectations and Language

As a first step in addressing the complicated issue of reliability for performance assessment data, assessment specialists need first to examine their own assumptions, expectations, and language when dealing with the notion of reliability. For instance, when assessment specialists talk with one another about reliability it most often centers on the reliability of a standardized achievement test and takes the form of a KR-20. However, internal consistency is often not the kind of reliability needed for quantitative observation data. In fact, the Standards for Educational and Psychological Tests (APA, 1985) recommend against the

use of any one reliability as the reliability of a particular assessment or test. Measures of inter- and, or intraobserver reliability are often more useful for performance assessments. I would also submit that the expectation of the magnitude of this reliability coefficient discussed by assessment specialists will be close to 0.90 (since most coefficients are for standardized achievement tests) and if they are not the dependability of the scores from these tests are often thought to be suspect. This standard must change when using performance assessments. By definition, the nature of a performance assessment demands that many of the variables that are fixed in standardized tests are left free to vary in the performance assessment. Consequently, the magnitude of the reliability coefficient may be less than would be obtained for a standardized test. The beauty and potential use of the performance assessment is to obtain outcome data on valued achievement targets difficult to assess with standardized achievement tests. The potential gain in further achievement data is obtained at a loss in the magnitude of the reliability coefficient. In effect, the performance assessment forces the examination of variance which was otherwise relegated to error and

"standardized out of the observed score variance " in the multiple-choice assessment format (Suen & Ary, 1989).

In addition to changing the expectation of the magnitude of the reliability coefficient, assessment specialists involved with large-scale performance assessments, must re-claim the word **reliability**. That is to say, reliability may refer to more types of reliability than internal consistency. The notion however, of many different types of reliability for one assessment is certainly not new. Several psychometricians have addressed this notion over the last several years (e.g., Cone, 1977; Cronbach, Gleser, Nanda, & Rajaratnam, 1962; Feldt & McKee, 1958). Nevertheless, we need to be reminded of this fact, again, given the pervasive nature of standardized tests, the kind of reliability reported for these tests, and the uniqueness of the performance assessment. Interrater reliability, intrarater reliability, reliability generalized to one rater, or generalized to the mean of several raters are examples of the types of reliability and subsequent parlance which should be an integral part of the measurement

specialists vocabulary when assessing the dependability of the scores from performance assessments.

A Dearth of Training, Resources, and Research

Assessment specialists in school districts and state agencies come to their positions with a wide range of measurement backgrounds. Some have been formally trained in measurement while others have had only a small number of measurement courses. Consequently, the assessment training needs of these professionals will vary. Whatever the level of training of the specialist, if their training was similar to mine, they received a quality education for the development and analysis of paper and pencil tests. However, only a cursory exposure to the psychometric issues of performance assessments was available. Certainly, a major reason for this is that graduate training often reflects the needs of educators. Standardized tests and other paper and pencil assessments dominated assessment practices during the time of my training. Consequently, the content of most measurement courses deal almost exclusively with the development and evaluation of paper and pencil tests to the exclusion of most other measurement topics

(Stiggins, 1990). The lack of training for assessment specialists concerning the psychometric issues of quantitative observation data may in part explain the current dearth of reliability data for performance assessments.

There is also a lack of available resources for assessment specialists regarding the psychometric issues of quantitative observation data. Most textbooks in educational measurement for example, are of high quality and provide excellent coverage for the development of standardized tests and other paper and pencil assessments. However, many of these texts do not provide coverage of other assessment practices, such as estimating reliability in large-scale performance assessments (Stiggins & Conklin, 1988).

Yet another reason which may in part explain the dearth of reliability data for performance assessments is the nature and focus of research by measurement specialists. for the most part, this research is focused on the development of paper and pencil tests (Stiggins, 1990). To explore this issue further as related to the reliability of quantitative observation data a check on the work of measurement researchers was conducted by examining the content of papers published

in the Journal of Educational Measurement (JEM). This was conducted by using the key word, **interrater reliability**, to determine the number of articles published on this topic in JEM from 1979 - present. Interrater reliability is most often used to describe reliability associated with quantitative observation data. The findings showed that of 292 articles published during this time, only four (1.4%) dealt with this topic.

Our education, resources, and research are of high quality, and focus on the development and analysis of standardized tests and other paper and pencil assessments. While this high quality and interest has served the education community well for many years, the growth in interest of performance assessments indicates that our training, resources, and research need to change.

The Reliability of Quantitative Observation Data

Few topics in the field of educational and psychological measurement are as misunderstood as the topic of measuring the reliability of quantitative observation data (Cone, 1977; Suen & Ary, 1989). Despite having appropriate techniques and methodology

available for decades, establishing proper reliability information for observation data continues to be illusive for many researchers. Hartmann (1981) argues that much of the problem lies with the discipline-specific nature of the literature on the topic, each with its own set of circumstances and language to describe what otherwise ought to be similar psychometric concerns and methodology. This xenophobia, as well as the lack of training available, have kept the field fractured and stymied. This state of affairs is no longer tolerable as performance assessments continue to gain prevelance in our schools.

Agreement Versus Reliability

Historically, the most popular method of measuring reliability has been to compute the percentage of times observers agree in their ratings of a particular phenomenon. Mitchell (1979) found that the majority of papers published in the journals, Child Developmental, and Developmental Psychology which used quantitative observation data reported reliability as percent agreement. There is intuitive appeal to this approach; certainly, we expect as a prerequisite to dependable scores agreement between the people declaring those

scores. Along with the ease of computing the percent agreement, the intuitive appeal may in part explain its prevelance. Despite the ease of computation and its intuitive appeal, there are limitations to this index which seriously detract from its usefulness. First, agreement is classified as an all or none phenomenon. Two or more people either agree or they don't. No measure of the degree of reliability can be obtained. Second, this index capitalizes on chance agreement. There is no correction for raters simply agreeing on a random basis, which can occur when the target behavior occurs with a high or low frequency. Suen and Ary (1989) eloquently illustrate the problem this way:

> Consider a situation in which two observers are observing a behavior that, in actuality, occurs in 5 of 100 intervals within a 100-interval observation session. The first observer correctly reports behavior occurence in 5 of the 100 intervals. The second observer was distracted and did not record any behavior occurence at all. In this case, both observers agreed that the behavior did not occur in 95% of the 100 intervals, whereas they disagreed on the occurence/nonoccurence of the behavior in the remaining 5 intervals. Hence, the interobserver agreement index would be equal to 0.95.
>
> Despite the fact that the second observer completely missed the occurences of behavior, the two observers appear to have a high level

> of consistency. This appearance of consistency is due to chance agreement. This result erroneously implies a high level of interobserver agreement. In fact, two observers can be observing two totally unrelated events at two different points in time and place and still, through chance, show a spuriously high value of percentage agreement. In general, the more the actual prevelance of behavior occurence approaches 100% or zero, the more percentage agreement is possibly inflated by chance agreement (Costello, 1973; Hartmann, 1977; Hopkins & Herman, 1977; Johnson & Bolstad, 1973; Mitchell, 1979).
>
> (pp. 107-108)

Third, this index does not consider the mean or variance of the ratings which is necessary when conceptualizing the reliability as a ratio of true score variance to observed score variance. Fourth, and perhaps most revealing, one can have high agreement and low reliability, and visa versa (Johnson & Bolstad, 1973; Suen & Ary, 1989; Tinsley & Weiss, 1975). This fact is overlooked by many researchers. Therefore, due to the aforementioned limitations, percent agreement is a poor index of reliability for quantitative observation data.

However, measures of agreement can be useful when analyzing quantitative observation data for two reasons. One, agreement is tied to reliability in a

similar but more complicated way as reliability is connected to validity. Tinsley and Weiss (1975) and Chapmans, Fyans, and Kerins (1984) advocate the reporting of agreement indices to provide the user of the assessment results with a context to judge the subsequent reliability measures. Second, the training and ongoing monitoring of observers is critical in performance assessment. To this end, agreement information between observers at different points on the rating scale or checklist may prove to be useful for diagnosing problems observers may have in their interpretation of particular standards or preventing observer drift from the standards over time. These differences may in part explain a low reliability coefficient and re-training may be necessary to rectify this problem and may then increase the dependability of the scores. The GED Testing Service uses this approach to monitor the de-centralized scoring of the essay component of the GED examination (Patience & Swartz, 1987). Scorers are used at different sites to score essays. Agreement indices are computed between scores within each site and between scores from a particular site and the GED Testing Service central office.

Another source of confusion has been the plethora of measures of agreement. Many of them are similar to the percent agreement index. The least controversial index is Cohen's Kappa. It is computed in the following way:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where, p_o is the observed proportion agreement and p_e is the expected agreement by chance. This measure is corrected for random agreement and can be thought of as the ratio of actual nonchance agreements to the total possible nonchance agreements. If measures of agreement are desired, Cohen's Kappa is the index of choice (Suen & Ary, 1989). For nominal data such as data obtained from a checklist, this statistic provides a measure of agreement and reliability (Berk, 1979; Tinsley & Weiss, 1975).

A Psychometric Theory for Quantitative Observation Data

Historically, much of the confusion and lack of coherence in the literature on the analysis of quantitative observation data is due the lack of a

psychometric theory for such data (Suen & Ary, 1989). The use of percent agreement as a measure of reliability for example, is not based on any psychometric theory. It is simply a statistical technique to analyze data.

The Random Sampling Theory and Item Response Theory are the two dominant theories in psychometrics today. Item response theory is in its infancy with regard to quantitative observation data and consequently, has little to offer as yet (Suen & Ary, 1989). The Random Sampling Model which assumes the existence of an infinite universe of possible subjects, items, observers, or observation points is currently the model of choice for quantitative observation data (Brennan, 1983; Suen & Ary, 1989). Suen and Ary (1989) have provided the first discussion of the Random Sampling Model as it pertains to the analysis of quantitative observation data and interested readers are referred to this source.

Intraclass Correlation Coefficient

There is growing concensus that the intraclass correlation coefficient (ICC) provides the best means for establishing reliability for quantitative

observation data (Armstrong, 1981; Berk, 1979; Feldt & McKee, 1958; Suen & Ary, 1989; Tinsley & Weiss, 1975). The ICC can be further couched within the rubric of generalizability theory. Generalizability Theory, which has been specifically advocated for use with large-scale performance assessments, is based on the Random Sampling psychometric model (Chapman, Fyans, and Kerins, 1984; Shavelson, Carey, & Webb, 1990).

Through the analysis of variance the ICC partitions the observed score variance into variance components to provide estimates of true score variance and error variance given the important variables in the study. These variance components are then combined to form a ratio which depicts the proportion of observed score variance which is true score variance (subject variance). The role of the ICC when determining interrater reliability, for example, is to capture the extent to which the ratings of different observers are proportional to one another when expressed as deviations from their means (Tinsley & Weiss, 1975). Since the ICC is not covered in most measurement texts, I will undertake a detailed discussion at this time.

The Models

There are three basic ICC models one can use depending on the design assumptions. Within each of these models is the ability to generalize to one observer or the mean of the observers in the study. For each model, the formula for generalizing to the mean of the observers in the study is obtained by applying the Spearman-Brown formula to the model which generalizes to one observer. I will discuss each of these models. The computational formulas for each model with the unit of analysis being one observer can be found in Table 1. The computational formulas for each model with the unit of analysis being the mean of the observers' ratings can be found in Table 2. Table 3 contains the corresponding theoretical formulas to Table 1. Table 4 contains the corresponding theoretical formulas to Table 2.

INSERT TABLE 1 ABOUT HERE

INSERT TABLE 2 ABOUT HERE

INSERT TABLE 3 ABOUT HERE

INSERT TABLE 4 ABOUT HERE

Model I: One-Way ANOVA

For some studies the task of assessing the performance of different subjects is assigned to different observers. Said another way, different subjects are nested within each observer. The correct way to compute the reliability for this design is to use the one-way ANOVA model. Notice that in this model residual variance is confounded with rater variance and is subtracted out of the numerator of the reliability coefficient and added to the denominator. This model has been recommended when observer variance is thought to be error (Armstrong, 1981; Bartko, 1976; Tinsley &

Weiss, 1975). This may occur when observers have been trained. High agreement between observers is expected after training, so that any difference between observers is thought to be error.

Model II: Two-Way Random Effects

When all observers have observed the same subjects and observer variance is thought to be error the two-way random effects model is appropriate ICC. This model treats the effects for observers and subjects as random. Writing assessment programs, for example, often use two observers to judge the quality of a group of papers. In this instance three variance components are obtained: a subjects variance component, an observer variance component and a residual variance component. The variance components for observers and residual can be separated in this model because all observers observe all subjects. The separation of the observer and residual variance in this model is in contrast with the one-way ANOVA model where observers observe different subjects and consequently, the observer and residual variance components are confounded. The reliability (generalizability) coefficient for the two-way random effects model has the variance component for subjects in the numerator

and the variance components for subjects, raters, and residual in the denominator.

Model III: Two-Way Mixed Model

This model treats the observer variable as fixed. In the numerator of the subsequent reliability coefficient is the subjects variance component. The denominator contains the subject and residual variance components. This model does not generalize to observers since this variable is fixed. It is used when generalization is desired to the observers in a particular study. The health sciences sometimes use this model (Armstrong, 1981). Also, this model offers a norm-referenced reliability estimate for performance assessments (Suen & Ary, 1989) which I will discuss in detail later in this paper.

Design Decisions

There are two key design decisions when using the aforementioned ICC models. The first is whether or not to generalize to one or the mean of several observers. Generalizing to the mean of a group of observers will often generate a higher reliability coefficient than generalizing to one observer. There is more

information in generalizing to the mean and this usually translates to an increase in the magnitude of the reliability coefficient. However, to claim that the dependability of the measures can be summarized by an ICC which generalizes to the mean of a group of observers means that whenever a particular performance is assessed the number of observers which constituted the mean will have to be deployed. This is often difficult logistically. Generalizing to one scorer may then be preferable. The implication is that the dependability of the scores will be maintained with one rater scoring the performance. There are few examples of this design in the education literature, however. Perhaps one reason is that it may be counterintuitive to expect one trained observer, no matter how good, to score in an unbiased manner. Therefore, two or more observers are often used to score a particular performance. There is perhaps some notion of accountability among the observers when using this procedure. Nevertheless, this is one issue that must be decided upon before a reliability study is implemented.

The second decision is whether or not to include observer variance as error variance. There are two

perspectives which can be used to make this decision: norm-referenced and criterion-referenced. Model III is most useful with norm-referencing as this model does not employ variance due to judges in the denominator of the ICC. Observer variance will not have an impact on the rank ordering of individuals. Model II is appropriate with criterion-referencing since observer bias is an issue. Consequently, differences between observers will detract from true score variance. This model has the observer variance component in the denominator of its ICC. Since model III does not have the observer variance component in the denominator of its ICC it will generally yield a higher reliability coefficient than Model II.

Another question is whether to use Model I or Model II if observer variance is not part of true score variance. Many advocate Model I (e.g., Armstrong, 1981; Shrout & Fleiss, 1979; Tinsley & Weiss, 1975), partly due to the formula for this model; observer variance is subtracted from the between persons mean square in the numerator of the reliability coefficient, as well as being added to the denominator. By adding observer variance to the denominator and subtracting it from the numerator, the variance is highlighted.

However, because this model has been lauded as the model of choice when observer variance is thought to be error, it is applied in situations where all observers rate all subjects, which is a design appropriate for Model II. Perhaps some clarity can be obtained by considering the design. If subjects are nested within observers, model I is the only possible approach. If all observers score all subjects Model I and Model II provides similar results when there are few differences between observers. Model I would be a conservative approach to estimating reliability since rater variance is confounded with residual variance. This combined variance is subtracted from the numerator which will provide a lower reliability coefficient than Model II if there are significant differences between observers.

An Example

A common design for performance assessments in education is to have all observers score all subjects. Consider the following data obtained from Winer (1962, p.127):

2 4 3 3

5 7 5 6

1 3 1 2

7 9 9 8

2 4 6 1

6 8 8 4

This matrix is 6 person by 4 observer matrix. Applying two-way ANOVA to this data results in the following analysis:

Source of variation	SS	df	MS
Subjects	122.50	5	24.50
Within subjects	36.00	18	2.00
Observers	17.50	3	5.83
Residual	18.50	15	1.23
Total	158.50	23	

Variance components

$\sigma^2_{subject} = (MS_{subject} - MSresid)/n_o = 7.76$

$\sigma^2_{observer} = (MS_{observer} - MS_{resid})/n_s = 0.92$

$\sigma^2_{residual} = MS_{resid} = 1.23$

Model III is appropriate from a norm-referencing perspective. The reliability coefficient is 0.83. Model II is appropriate if criterion-referenced perspective is desired. The reliability coefficient is 0.74.

Analysis of Variance Components

Many have advocated the use of analyzing the magnitude of the variance components as a preferrable and more informative analysis than simply looking at the reliability coefficient (e.g., Cronbach, et. al, 1972; Brennan, 1983; Suen & Ary, 1989). In fact, supplying estimates of variance components for a particular assessment is recommended in the educational and psychological test standards (APA, 1985). Examination of the variance components in the above example shows a fairly small value for the observer variance component at 0.92. This is desirable since we do not want observers to differ from one another. The subject (true score) variance is high at 7.76. It is also high compared to the observer variance component, which is also desirable. From a norm-referenced perspective a large subjects variance component means that differences between subjects may be more easily

determined than if the variance component were small. From either a norm-referenced or criterion-referenced perspective a large subjects variance component is desirable since it is in the numerator of the reliability coefficient. As the numerator of the ICC ratio increases in magnitude compared with the denominator, the overall reliability increases.

Available Computer Software

I have anectodal evidence to suggest that even when researchers know to use the intraclass correlation for a specific research problem, often what is computed and reported is the coefficient available in a given software package. This ICC may or may not be appropriate for the research question at hand. Therefore, part of the problem in establishing reliability for quantitative observation data is to know the capabilities of the software one is using and to be able to interpret the results.

Any computer program which computes ANOVA will provide the necessary information for the reliability of the data. Some of these programs go a step further and combine the mean squares from the ANOVA table to compute variance components. One simply then needs to

combine the appropriate variance components in a specific way to obtain the intraclass correlation coefficient.

SPSSX provides the mean squares from a matrix of quantitative observation data in its reliability module. The user must combine these mean squares however, to obtain any of the ICCs. Their offices are located in Chicago, Illinois.

Baumgartner (1987) has developed a program for the Apple II personal computer which computes the intraclass correlation. This can be obtained by writing to the author at: Department of Physical Education, University of Georgia.

Paulson and Trevisan (1990) have developed an MS-DOS program which computes the intraclass correlation and can be obtained by writing to: F. L. Paulson, Multnomah ESD, 11611 NE Ainsworth Circle, Portland, Oregon 97220-1039. Send a self-addressed stamped mailer with a 5 1/4" or 3 1/2" diskette.

Discussion and Conclusion

This paper identified historical and training issues which beset the proper psychometric analysis of quantitative observation data from performance

assessments. There has never been a better time however, to correct this situation. Performance assessments are rapidly increasing in number. Little reliability data is available on these measures.

The challenge facing the measurement community is to make sure these assessments are done reliably and communicate this information in an understandable manner to decisionmakers using performance assessment data. To do this, many of us will need to learn techniques which may not have been part of our training. Accessible resources will need to be written for the practitioner. We also need research focused on the psychometric issues of performance assessments.

References

American Psychological Association (1985). Standards for educational and psychological testing. Washington, D.C.

Armstrong, Gordon D. (1981, October). The intraclass correlation as a measure of interrater reliability of subjective judgements. Nursing Research, pp. 314-315, 320A.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765.

Baumgartner, T. A. (1989). Norm-referenced Measurement: Reliability. In M. J. Safrit & T. Wood (Eds.) Measurement Concepts in Physical Education and Exercise Science (pp. 45-72). Champaign, IL: Human Kinetic Books

Berk, Ronald A. (1979). Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. American Journal of Mental Deficiency, 83, 5, 460-472.

Brennan, R. L. (1983). Elements of Generalizability Theory. Iowa City: ACT Publications.

Chapman, C. W., Fyans, Jr., L. J., and Kerins, C. T. (1984). Writing Assessment in Illinois. Educational Measurement: Issues and Practices, 3, 1, 24-26.

Cone, J. D. (1977). The relevance of reliability and validity for behavior assessment. Behavior Therapy, 8, 411-426.

Cronbach, L. J., Gleser, G. C., Nanda, H., Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley and Sons.

Feldt, L. S. & McKee, M. E. (1958). Estimation of the reliability of skill tests. Research Quarterly, 29, 3, 279-293.

Haggard, Ernest A. (1958). Intraclass correlation and the analysis of variance. New York: The Dryden Press, Inc.

Hartmann, D. P. (1981). Editorial note. Behavioral Assessment, 3, 1-3.

Johnson, S. M. & Bolstad, O. D. (1973). Methodological Issues in Naturalistic Observations: Some Problems and Solutions for Field Research. In L. A. Hamerlynck, L.C. Hardy, and E. J. Mash (Eds.), Behavior Change: Methodology, Concepts, and Practice (pp. 7-67). Champaign, IL: Research Press.

Mitchell, S. K. (1979). Interobserver agreement, reliability and generalizability of data collected in observational studies. Psychological Bulletin, 86, 2, 376-390.

Patience, W. & Swartz, R. (1987, April). Essay score reliability: issues in and methods of reporting GED writing skills test scores. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

Paulson, F. L. & Trevisan, M. S. (1990). INTRACLS: application of the intraclass correlation to computing reliability [Computer program]. Applied Psychological Measurement, 14, 2, 212.

Rothman, R. (1990, September 12). New tests based on performance raise questions. Education Week, pp. 1, 10, 12.

Shavelson, R. J., Carey, N. B., and Webb, N. M. (1990, May). Indicators of science achievement: options for a powerful policy instrument. Phi Delta Kappan, 692-697.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 2, 420-428.

Stiggins, R. J. (In press). Assessment Literacy. Phi Delta Kappan.

Stiggins, R. J. & Conklin, N. F. (1988). Teacher training in assessment. Portland, OR: Northwest Regional Educational Laboratory.

Suen, H. K. & Ary, D. (1989). Analyzing Quantitative Behavioral Observation Data. New Jersey: Lawrence Erlbaum Associates, Inc.

Suen, H. K. & Davey B. (1990, April). Potential theoretical and practical pitfalls and cautions of the performance assessment design. Symposium paper presented at the Annual Meeting of the American Educational Research Association, Boston, MA.

Suen, H. K. (1991). Personal communication, February 5.

Tinsley, H. A. & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. Journal of Counseling Psychology, 22, 4, 358-376.

Winer, B. J. (1962). Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company.

Table 1. Formulas for the Intraclass Correlation Individual Observer Unit of Analysis

$$ICC_1 = \frac{BMS-WMS}{BMS+(K-1)WMS}$$

Generalize from different observers; 1-way ANOVA for subjects
Model I

$$ICC_2 = \frac{BMS-RMS}{BMS+(K-1)RMS+K(OMS-RMS)/N}$$

Generalize from same observers; 2-way ANOVA subjects by observers; Model II

$$ICC_3 = \frac{BMS-RMS}{BMS+(K-1)RMS}$$

No Generalization; observer variable is fixed; 2-way ANOVA subjects by observers; Model III

Where, BMS = between persons mean square
WMS = within person mean square
OMS = observer mean square
RMS = residual mean square
K = number of observers

Table 2. Formulas for the Intraclass Correlation
Mean of the Observers' Ratings
Unit of Analysis

$$ICC_4 = \frac{BMS-WMS}{BMS}$$

Generalize from different observers; 1-way ANOVA for subjects
Model I

$$ICC_5 = \frac{BMS-RMS}{BMS+(OMS-RMS)/N}$$

Generalize from same observers; 2-way ANOVA subjects by observers; Model II

$$ICC_6 = \frac{BMS-RMS}{BMS}$$

No generalization; observer variable is fixed; 2-way ANOVA subjects by observers; Model III

Where, BMS = between persons mean square
WMS = within persons mean square
OMS = observer mean square
RMS = residual mean square
K = number of observers

Table 3. Theoretical Formulas for the Intraclass Correlation (Single Observer Unit of Analysis)

$$ICC_1 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_w}$$

$$ICC_2 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_o + \sigma^2_r}$$

$$ICC_3 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_r}$$

Where, σ^2_s = estimated subjects variance component

σ^2_w = estimated within subjects variance component

σ^2_o = estimated observer variance component

σ^2_r = estimated residual variance component

Table 4. Theoretical Formulas for the Intraclass Correlation (Mean Observers' Ratings Unit of Analysis)

$$ICC_4 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_w/K}$$

$$ICC_5 = \frac{\sigma^2_s}{\sigma^2_s + (\sigma^2_o + \sigma^2_r)K}$$

$$ICC_6 = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_r/K}$$

Where K = Number of observers